The results of biodiversity–ecosystem functioning experiments are realistic

Jochum, Malte ; Fischer, Markus ; Isbell, Forest ; Roscher, Christiane ; van der Plas, Fons ; Boch, Steffen ; Boenisch, Gerhard ; Buchmann, Nina ; Catford, Jane A ; Cavender-Bares, Jeannine ; Ebeling, Anne ; Eisenhauer, Nico ; Gleixner, Gerd ; Hölzel, Norbert ; Kattge, Jens ; Klaus, Valentin H ; Kleinebecker, Till ; Lange, Markus ; Le Provost, Gaëtane ; Meyer, Sebastian T ; Molina-Venegas, Rafael ; Mommer, Liesje ; Oelmann, Yvonne ; Penone, Caterina ; Prati, Daniel ; Reich, Peter B ; Rindisbacher, Abiel ; Schäfer, Deborah ; Scheu, Stefan ; Schmid, Bernhard ; et al

Abstract: A large body of research shows that biodiversity loss can reduce ecosystem functioning. However, much of the evidence for this relationship is drawn from biodiversity–ecosystem functioning experiments in which biodiversity loss is simulated by randomly assembling communities of varying species diversity, and ecosystem functions are measured. This random assembly has led some ecologists to question the relevance of biodiversity experiments to real-world ecosystems, where community assembly or disassembly may be non-random and influenced by external drivers, such as climate, soil conditions or land use. Here, we compare data from real-world grassland plant communities with data from two of the largest and longest-running grassland biodiversity experiments (the Jena Experiment in Germany and BioDIV in the United States) in terms of their taxonomic, functional and phylogenetic diversity and functional-trait composition. We found that plant communities of biodiversity experiments cover almost all of the multivariate variation of the real-world communities, while also containing community types that are not currently observed in the real world. Moreover, they have greater variance in their compositional features than their real-world counterparts. We then re-analysed a subset of experimental data that included only ecologically realistic communities (that is, those comparable to real-world communities). For 10 out of 12 biodiversity–ecosystem functioning relationships, biodiversity effects did not differ significantly between the full dataset of biodiversity experiments and the ecologically realistic subset of experimental communities. Although we do not provide direct evidence for strong or consistent biodiversity–ecosystem functioning relationships in real-world communities, our results demonstrate that the results of biodiversity experiments are largely insensitive to the exclusion of unrealistic communities and that the conclusions drawn from biodiversity experiments are generally robust.

DOI: https://doi.org/10.1038/s41559-020-1280-9

Markus; Le Provost, Gaëtane; Meyer, Sebastian T; Molina-Venegas, Rafael; Mommer, Liesje; Oelmann, Yvonne; Penone, Caterina; Prati, Daniel; Reich, Peter B; Rindisbacher, Abiel; Schäfer, Deborah; Scheu, Stefan; Schmid, Bernhard; et al (2020). The results of biodiversity–ecosystem functioning experiments are realistic. Nature Ecology and Evolution, 4(11):1485-1494.
DOI: https://doi.org/10.1038/s41559-020-1280-9

Title:

The results of biodiversity-ecosystem functioning experiments are realistic

Running headline:

Realism of biodiversity experiments

Authors:

Jochum, Malte[1, 2, 3], Fischer, Markus[1], Isbell, Forest[4], Roscher, Christiane[5, 2], van der Plas, Fons[6], Boch, Steffen[1, 7], Boenisch, Gerhard[8], Buchmann, Nina[9], Catford, Jane A.[10], Cavender-Bares, Jeannine[4], Ebeling, Anne[11], Eisenhauer, Nico[2, 3], Gleixner, Gerd[8], Hölzel, Norbert[12], Kattge, Jens[2, 8], Klaus, Valentin H.[9], Kleinebecker, Till[13], Lange, Markus[8], Le Provost, Gaëtane[14], Meyer, Sebastian T.[15], Molina-Venegas, Rafael[1, 16], Mommer, Liesje[17], Oelmann, Yvonne[18], Penone, Caterina[1], Prati, Daniel[1], Reich, Peter B.[19, 20], Rindisbacher, Abiel[1], Schäfer, Deborah[1], Scheu, Stefan[21, 22], Schmid, Bernhard[23, 24], Tilman, David[25, 26], Tscharntke, Teja[27], Vogel, Anja[2, 3, 11], Wagg, Cameron[28], Weigelt, Alexandra[2, 29], Weisser, Wolfgang W.[15], Wilcke, Wolfgang[30], Manning, Peter[14]

Affiliations:

1. Institute of Plant Sciences, University of Bern, Altenbergrain 21, 3013 Bern, Switzerland

2. German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany

3. Leipzig University, Institute of Biology, Deutscher Platz 5e, 04103 Leipzig, Germany

4. Department of Ecology, Evolution and Behavior, University of Minnesota, 1479 Gortner Ave, Saint Paul MN 55108, USA

5. UFZ, Helmholtz Centre for Environmental Research, Department Physiological Diversity, Permoserstrasse 15, 04318 Leipzig, Germany
6. Systematic Botany and Functional Biodiversity, Institute of Biology, Leipzig University, Johannisallee 21-23, 04103 Leipzig, Germany
7. WSL Swiss Federal Research Institute, Zürcherstr. 111, 8903 Birmensdorf, Switzerland
8. Max Planck Institute for Biogeochemistry, Hans-Knoell Str. 10, 07745 Jena, Germany
9. Institute for Agricultural Sciences, ETH Zürich, Universitätstr. 2, 8092 Zürich, Switzerland
10. Department of Geography, King's College London, 30 Aldwych, London, WC2B 4BG, UK
11. Institute of Ecology and Evolution, University of Jena, Dornburger Str. 159, 07743 Jena, Germany
12. Institute of Landscape Ecology, University of Münster, Heisenbergstr. 2, 48149 Münster, Germany
13. Institute of Landscape Ecology and Resources Management, Giessen University, Heinrich-Buff-Ring 26, 35392 Giessen, Germany
14. Senckenberg Biodiversity and Climate Research Centre (SBiK-F), Senckenberganlage 25, 60325 Frankfurt, Germany
15. Terrestrial Ecology Research Group, School of Life Sciences Weihenstephan, Technical University of Munich, Hans-Carl-von-Carlowitz-Platz 2, 85354 Freising, Germany
16. GloCEE--Global Change Ecology & Evolution group, Dept. Life Sciences, University of Alcalá, 28805 Alcalá de Henares, Spain
17. Plant Ecology and Nature Conservation Group, Dept. of Environmental Sciences, Wageningen University, PO box 47; 6700 AA, Wageningen, the Netherlands
18. Geoecology, Geoscience Department, University of Tübingen, Rümelinstr. 19-23, 72070 Tübingen, Germany
19. Department of Forest Resources, University of Minnesota, 1530 Cleveland Avenue N, St. Paul, MN 55108, USA
20. Hawkesbury Institute for the Environment, Western Sydney University, Penrith South DC, NSW 2751, Australia

21. JF Blumenbach Institute of Zoology and Anthropology, University of Goettingen, Untere Karspüle 2, 37073 Göttingen, Germany
22. Centre for Biodiversity and Sustainable Land Use, Von-Siebold Str. 8, University of Goettingen, 37075 Göttingen, Germany
23. Department of Geography, University of Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland
24. Institute of Ecology, College of Urban and Environmental Sciences, Peking University, 100871 Beijing, China
25. Department of Ecology, Evolution, and Behavior, University of Minnesota, St. Paul, MN, 55108, USA
26. Bren School of Environmental Science and Management, University of California, Santa Barbara, CA, 93106, USA
27. Agroecology, Dept. of Crop Sciences, University of Göttingen, Grisebachstrasse 6, 37077 Göttingen
28. Department of Evolutionary Biology and Environmental Studies, University of Zürich, Winterthurerstrasse 190, Zürich CH-8047, Switzerland
29. Leipzig University, Institute of Biology, Johannisallee 21-23, 04103 Leipzig, Germany
30. Karlsruhe Institute of Technology (KIT), Institute of Geography and Geoecology, Reinhard-Baumeister-Platz 1, 76131 Karlsruhe, Germany

Materials and Correspondence: Please address requests to **malte.jochum@idiv.de**.

Summary

A large body of research shows that biodiversity loss can reduce ecosystem functioning, thus providing support for the conservation of biological diversity[1–4]. Much of the evidence for this relationship is drawn from biodiversity-ecosystem functioning experiments (hereafter: biodiversity experiments), in which biodiversity loss is simulated by randomly assembling communities of varying species diversity, and ecosystem functions are measured[5–9]. This random assembly has led some ecologists to question the relevance of biodiversity experiments to real-world ecosystems, where community assembly or disassembly may be non-random and influenced by external drivers, such as climate, soil conditions or land use[10–19]. Despite these repeated criticisms, there has been no comprehensive, quantitative assessment of how experimental and real-world plant communities really differ, and whether these differences invalidate the extrapolation of experimental results to natural systems. Here, we compare data from two of the largest and longest-running grassland biodiversity experiments (Jena Experiment, Germany; BioDIV, USA) to related real-world grassland plant communities in terms of their taxonomic, functional, and phylogenetic diversity and functional-trait composition. We found that plant communities of biodiversity experiments cover almost all of the multivariate variation of the real-world communities, while also containing community types that are not currently observed in the real world. Moreover, they have greater variance in their compositional features than their real-world counterparts. We then re-analysed a subset of experimental data that included only ecologically-realistic communities, i.e. those comparable to real-world communities. For ten out of twelve biodiversity-ecosystem functioning relationships, biodiversity effects did not differ significantly between the full dataset of biodiversity experiments and the ecologically-realistic subset of experimental communities. Although we do not provide direct evidence for strong or consistent biodiversity-ecosystem functioning relationships in real-world communities, our results demonstrate that the results of biodiversity

86 experiments are largely insensitive to the exclusion of unrealistic communities. By bridging the gap
87 between experimental and real-world studies, this study shows that the conclusions drawn from
88 biodiversity experiments are generally robust, a key step in translating their results into the context of
89 real-world ecosystems.

Main Text

Concerns over the consequences of biodiversity loss for human well-being triggered the growth of biodiversity-ecosystem functioning (hereafter: biodiversity-functioning) research, an important field of ecology over the past 25 years[1,3,20–23]. Some of the most influential studies in this field are based on biodiversity-ecosystem functioning experiments (hereafter: biodiversity experiments), in which communities of varying diversity are randomly assembled and the responses of ecosystem processes are measured[6,24]. These experiments, often conducted using grassland communities[8], aim to isolate the effects of species richness from other factors known to affect ecosystem processes, such as climate, nutrient availability, and the presence of particular plant functional types. By doing so, they have provided strong evidence that biodiversity can affect the functioning of ecosystems – most commonly with a positive but saturating relationship between diversity and plant productivity[1,2,5,7,22,25,26]. However, the relevance of biodiversity experiments to real-world ecosystems (i.e., those where community assembly is influenced by external drivers, such as climate, soil conditions or land use) has been repeatedly questioned[10–14,18]. Criticisms highlight several common features of experimental designs, namely random assembly, as opposed to non-random assembly/disassembly of real-world ecosystems[13], initial sowing of even species abundances (but see[27–30]), and the repeated removal of non-target species (but see[31,32]). These factors may alter community assembly processes, leading to unrealistic communities that possess functional properties that are rare or absent in the real world. Although numerous researchers have argued for the relevance of biodiversity experiments[15,16,33,34] and provided evidence to counter these criticisms[28,35,36], we do not know how closely plant communities in biodiversity experiments resemble those of related real-world ecosystems (but see[37] for a local-scale comparison), or if the presence of unrealistic communities affects the conclusions drawn from these experiments.

To close these knowledge gaps, we take a two-step approach: first, we perform a comprehensive, quantitative assessment of the differences and similarities between plant communities from biodiversity experiments and related real-world ecosystems. Second, we test the robustness of conclusions drawn from biodiversity experiments to the removal of "unrealistic" communities - those least comparable to real-world communities. In the first step, we quantitatively compared the plant communities of two of the World's largest and longest-running grassland biodiversity experiments to those of nearby real-world communities where diversity gradients are created by natural environmental variation and global-change drivers. These experiments are the Jena Experiment, established 2002 in Jena, Germany (hereafter: Jena Experiment)[6,32] and the BioDIV experiment, established 1994 at the Cedar Creek Ecosystem Science Reserve, Minnesota, USA (hereafter: BioDIV)[5,38–40] (**Fig. 1**). We compared experimental communities from the Jena Experiment with those of agricultural grasslands in three regions of Germany, spanning a broad range of site conditions and land-use intensities – the Biodiversity Exploratories[41,42] – and grasslands close to the Jena Experiment (hereafter: "Jena real world"). BioDIV's experimental communities were compared to nearby, naturally-assembled prairie-grassland communities at Cedar Creek, including fertilized grasslands[35,43,44] and those undergoing successional change[45] (see Methods and Supporting Information, Table S1). We combined species-specific cover data from annual vegetation surveys (3,329 and 9,954 plot-year combinations in the German and the US datasets, respectively) with phylogenetic information and plant functional-trait data to characterize and quantitatively compare plant communities based on a range of properties known to represent important dimensions of biodiversity and to independently influence ecosystem functioning[46–49] including measures of taxonomic diversity and evenness, phylogenetic diversity, functional diversity and community abundance-weighted means (CWM) of selected functional traits of vascular plants, hereafter referred to as "community properties" in a Principal Component Analysis (PCA) (see

Methods for definitions of all community properties; **Fig. 1**). Based on this multidimensional, multivariate comparison of plant community properties, we identified plots from biodiversity experiments whose communities fell outside the multidimensional community-property space occupied by real-world plant communities (hereafter: "unrealistic communities"). This was achieved by calculating the intersection of three-dimensional convex hull volumes defined by experimental and real-world communities (**Fig 1**; see Methods). In the second step of our analysis, we fit linear models to test how plant species richness affected eight selected ecosystem functions from both the above- and belowground subsystems. This was done for both the full datasets and the subsets of realistic plots.

Results and discussion

Plant communities in biodiversity experiments and related real-world systems

The results of our multidimensional, multivariate comparison showed that experimental plant communities occupy a larger area of multivariate community-property space than real-world communities, despite the latter covering a wide range of climatic, edaphic and management conditions, particularly in the German dataset[41,50] (**Fig. 1a,e**). This finding was robust to the inclusion or exclusion of particular community properties and the choice of overlap calculation methodology (Supporting Information on sensitivity analyses I, Fig. S1 and Table S2, S3, S4) and was supported by additional data collected at Jena. This showed that experimental communities migrated towards the narrow space occupied by real-world communities when not weeded (i.e., Jena invasion, Supporting Information Fig. S2), thus also indicating that the differences between real-world and biodiversity-experiment communities in multivariate community-property space were due to experimental maintenance rather than differences in plot conditions, species pools or initially random versus natural community assembly.

158 Next, for each community property in each region (Germany and USA), we determined the
159 proportion of biodiversity-experiment plots that fell within the community-property range of the related
160 real-world plots (Supporting Information Fig. S3 and S4 and Tables S5 and S6). Specifically, in
161 Germany, SEve, S, PD, FRic, and MNTD showed the lowest proportion of biodiversity-experiment
162 plots in the real-world range of these properties. Experimental communities at Jena showed higher
163 values of SEve and MNTD and lower S, PD and FRic than their real-world counterparts. In contrast, in
164 the US dataset, it was LDMC, FEve, SLA, leaf N, and FRic that showed the lowest proportion of
165 experimental plots in the real-world range of community properties and all these community properties
166 showed lower values in the experimental than in the real-world communities.
167 Overall, three conclusions can be drawn from this comparative analysis: first, biodiversity
168 experiments successfully create plant communities that vary greatly in functionally-important
169 community properties. Second, real-world communities are confined to narrower regions of
170 multivariate community-property space than those of experiments. Third, while the properties of many
171 experimental communities are not observed in related real-world communities, a subset of randomly-
172 assembled experimental communities are functionally comparable to real-world communities, (**Fig. 1**
173 and Supporting Information Tables S5 and S6), even though their taxonomic community composition
174 may differ (see Supporting Information on sensitivity analyses I, section E, and Figure S5).
175 The comparative analysis was used to define which plant communities from biodiversity
176 experiments could be deemed comparable to real-world systems (hereafter 'realistic'). This revealed
177 that, when using 12 community properties selected using variance inflation factors (hereafter: vif) to
178 reduce redundant information (see Methods), 28% and 77% of experimental plots were deemed
179 realistic in Jena and BioDIV, respectively (Supporting Information Tables S3 and S4). The plant
180 communities of these realistic biodiversity-experiment plots had significantly higher sown diversity

(Jena: av = 21.7 realistic vs. 3.5 unrealistic, BioDIV: 7.8 vs. 1.7) and more sown functional groups (Jena: 2.8 vs. 1.9, BioDIV: 3.5 vs. 1.5), but lower Simpson's evenness (Jena: 0.5 vs. 0.7, BioDIV: 0.6 vs. 0.9; **Fig. 1**) than the unrealistic experimental plots (see **Fig. 1,** Supporting Information, Fig. S3 and S4, Table S7 and S8). Although the constraining was not based on species richness, the diversity gradient in Jena was truncated in the realistic subset of plots. In Jena, average minimum species richness across years was 1 in unconstrained (all plots) and 3.7 in the constrained datasets (realistic plots only). In contrast, BioDIV covered a relatively narrow range of species richness and the equivalent real-world communities were also relatively species poor, so here the gradient was not truncated (**Fig. 2** and Supporting Information, Table S11). As such, the low diversity plots in the Jena Experiment, although necessary for an experimental design that can identify diversity effects and their underlying mechanisms[51], are generally found to be unrealistic when compared to current German real-world communities. Note that study-specific differences in vegetation survey area could not be avoided, although their impact on the results was minimized (see methods and e.g. Supporting Information Fig. S9 for more detail).

The selection of realistic experimental plots was largely insensitive to most methodological choices, such as the exclusion of certain community properties and the overlap calculation method used (see Supporting Information on sensitivity analyses I for details). For example, using all 21 instead of only the 12 vif-selected community properties resulted in slight changes in the number and identity of plots selected as realistic (91-96 % of the main analysis plots included for Jena, 85-95 % for BioDIV; Tables S3 and S4). However, the selection of realistic plots was sensitive to some methodological choices. Within our sensitivity analyses, results were relatively sensitive to the following: changing the number of PCA axes used to compute multidimensional overlap, altering the criterion for defining inclusion / exclusion in the overlap, basing our comparison on species abundances rather than

community properties, and reducing the real-world data to include only those plots with comparable land use to the experiment (for details, see sections B, D, E, and F of Supporting Information on sensitivity analyses I, Tables S3 and S4, and Fig. S1 and S6). For example, when using species-abundance based NMDS to define realistic communities (Supporting Information Fig. S5), in the German dataset, very few experimental plots (2 %) fell within the real-world NMDS realm and were selected as realistic. For the USA dataset, 33 % of plots were selected as realistic. For BEF relationships based on these alternative analyses, see below. As such, as long as the overall analysis framework of using plant-community properties in PCAs to determine multidimensional overlap is used, as opposed to species-abundance based NMDS, our conclusions are robust to the methodological decisions taken.

Biodiversity-functioning relationships in unconstrained versus constrained experimental data subsets

Our comparison of biodiversity-functioning relationships in full datasets of biodiversity experiments (unconstrained, all plots) versus realistic subsets of plots (constrained, realistic plots only) was conducted for the following ecosystem functions: plant aboveground and belowground (root) biomass, plant aboveground carbon to nitrogen (C:N) ratio, soil organic carbon content, invertebrate leaf herbivory, soil microbial biomass C, phosphatase activity in the soil and pollinator abundance (**Fig. 2**). It showed that, in both experiments, and across the different ecosystem functions, the slopes of experimental biodiversity-functioning relationships were relatively insensitive to the removal of unrealistic communities (but see the discussion of significance changes below). A paired t-test on pairs of unconstrained and constrained slopes for the 12 BEF relationships shown in **Fig. 2** showed no significant change in slope estimates ($t=1.40$, $df=11$, $p=0.19$, $n=12$) and confidence intervals for slope

226 estimates overlapped each other's mean for all but two model pairs. The two exceptions to this were
227 both initially weak biodiversity-functioning relationships: Jena-Experiment herbivory, where the
228 positive slope increased when constrained to realistic plots, and BioDIV plant C:N, where a non-
229 significant, slightly negative slope turned into a positive significant one (see Supporting Information
230 Table S9). The finding that the slope of the biodiversity-functioning relationship was largely unaffected
231 by the exclusion of unrealistic communities was robust to changing the set of community properties in
232 the PCA and the method used to identify realistic communities (Supporting Information, Fig. S6). The
233 goodness of fit (adjusted R^2 values) was also only partly affected by constraining the dataset (mean R^2:
234 0.24 versus 0.15 for unconstrained and constrained models, respectively; Supporting Information, Table
235 S9), and the average percentage change in maximum functioning was ±10.3% (SE: 4%; Supporting
236 Information, Table S10). When using the realistic plots defined using all 21 instead of the 12 vif-
237 selected community properties in the PCA, BEF-slope changes from unconstrained to constrained data
238 subsets were largely unchanged (Supporting Information on sensitivity analyses I, Fig. S6). For
239 BioDIV, when using species-abundance based NMDS to define the overlap, constrained BEF
240 relationships were comparable to or more strongly positive than unconstrained relationships
241 (Supporting Information Fig. S6). Together, these results show that the form, strength, and magnitude
242 of the relationship between biodiversity and functioning that has been identified in biodiversity
243 experiments weakens somewhat, but is generally robust to the removal of unrealistic communities.
244 In four out of twelve cases, constraining data led to a change from a significant to a non-
245 significant relationship (Jena soil organic C content, root biomass, soil microbial biomass C and
246 phosphatase activity; **Fig. 2**). To check whether this change in significance was driven by the smaller
247 sample size of the constrained data set, we assessed the sensitivity of the results to reduced replication.
248 This was done by performing a sensitivity analysis in which we randomly reduced the size of the

unconstrained dataset of the Jena Experiment. This showed that the slope of the biodiversity-functioning relationship in the realistic subset for these four relationships was shallower than most slopes estimated from randomly selected data (Supporting Information on sensitivity analyses II, Fig. S7a). This suggests for certain ecosystem functions, particularly soil processes in the Jena Experiment, that the strength of the biodiversity-functioning relationship might be overestimated in biodiversity experiments.

The truncated species-richness gradient of the realistic plots at Jena was associated with a 31 % reduction in the range of functioning covered across the truncated reduced biodiversity gradient (Supporting Information, Table S11). Therefore, to investigate whether the shallower slope and loss of significance in realistic data subsets at Jena was driven by the truncation of the species-richness gradient, we performed an additional sensitivity analysis for the four Jena soil functions in question (Supporting Information on sensitivity analyses II and Fig. S7b). When we restricted the random choice of Jena Experiment plots to the shorter gradient of species richness covered by the realistic plots in the main analysis, the vast majority of BEF relationships in the sensitivity analysis turned non-significant (between 84 and 100 of 100 repetitions, see Fig. S7b). This indicates that it is primarily the shortened species-richness gradient, rather than reduced sample size, that drives the weakening of some BEF relationships when constrained (Supporting Information Fig. S7b). These results show that removing the lower end of the species-richness gradient leaves only the saturating, right-hand side of the commonly observed biodiversity-functioning relationship[1] in some constrained experimental datasets, for which the slope is shallower. These shallower slopes do not demonstrate that experiments falsely predict a stronger diversity-functioning relationship at low richness, but do indicate that some real-world systems do not vary over the full richness gradient found in experiments, thus potentially explaining the relatively weaker diversity-functioning relationships observed across real world

diversity gradients, compared to experiments[4].

Finally, several observational 'real-world' studies have shown that other aspects of biodiversity, e.g. functional composition, are stronger predictors of ecosystem functioning than species richness[4,52,53], while experiments show a dominant effect of species richness and related variables[47,54]. Therefore, we investigated whether the identity of the community properties that best explain function was affected by our constraining of ecosystem function. This demonstrated that the relative importance of plant-community properties in explaining experimental ecosystem functioning changed slightly due to the constraining to realistic experimental plots, with PD and CWM SLA gaining and FEve and MNTD losing importance for Jena aboveground biomass and SEve slightly gaining with CWM seed mass losing importance for BioDIV soil organic C (see Supporting Table S12 for details). However, there was no large systematic shift in the identity of the plant-community properties which best explain ecosystem functioning.

Our results show that the biodiversity-functioning relationships observed in biodiversity experiments are not an experimental artefact caused by the presence of unrealistic communities. The question remains, however, as to how important biodiversity is as a driver of ecosystem functioning in the real world, relative to factors such as land use or climate[7,14,55]. Although strong and positive biodiversity-functioning relationships have been reported in real-world studies[4,24,36,56–58], other studies describe weak or negative relationships[4,59,60]. This inconsistency, and the discrepancy between experimental and real-world patterns, is commonly attributed to the presence of covarying environmental or biological factors that also drive ecosystem functioning[61], and which obscure, confound or negate the effects of biodiversity (e.g., nutrient availability, climate, and the dominant functional traits of the community;[52,56,62–64]). These factors are likely to be closely coupled in real-world ecosystems, but decoupled in experiments. Indeed, across our datasets, the average correlation strength

295 of the eight measures of dominant functional traits (CWM's) with Simpson's evenness, functional, and
296 phylogenetic diversity properties was slightly higher in real-world than in experimental data subsets;
297 mean absolute correlation coefficients were 0.18 and 0.22 in German and US real-world plots,
298 compared to 0.08 and 0.16 in their respective experiments (Supporting Information, Table S13 and
299 S14).

300 While it would be desirable to directly compare the experimental biodiversity-functioning
301 relationships described in this study to those observed in real-world systems, both theoretical and
302 empirical studies show that simple, bivariate relationships between species richness and functioning
303 will not necessarily be positive, even if there are strong underlying effects of biodiversity on ecosystem
304 functioning[56,62]. For the main real-world datasets included in our study, previous investigations have
305 shown neutral or negative relationships between plant species richness and biomass for the German real
306 world dataset[65]. Furthermore, the relationship between species richness and a "production-only"
307 ecosystem-service scenario, heavily based on plant shoot biomass, was negative, even when accounting
308 for land-use intensity in a structural equation modeling framework[52]. This negative relationship may be
309 driven by extremely strong covariation between species richness and functional composition (species
310 richness - CWM SLA Pearson correlation is as strong as r=-0.9 in one region), making it virtually
311 impossible to distinguish between the effects of diversity and functional composition using
312 conventional methods. For the fertilization studies at Cedar Creek, negative relationships between
313 diversity and productivity across space were observed because fertilized plots possess high productivity
314 and low diversity, but when fertilization reduced plant species richness, this also reduced productivity
315 over time[35]. Consequently, adequately investigating real-world biodiversity-functioning relationships
316 requires specific, in-depth knowledge of the identity and interplay of additional drivers of both species
317 richness and ecosystem functions[56,62] and analysis frameworks capable of disentangling covariation in

318 and simultaneous reciprocal effects between these interrelated drivers.

319 While the biodiversity experiments used in our analysis cover a wide range of plant-community
320 properties, only a fraction of this multidimensional space is occupied by related real-world
321 communities. The remainder of space covered by the experimental communities is currently not
322 observed in the real-world communities that we considered; however, this "unrealized plant community
323 property space" may be useful in predicting ecosystem functioning in the future, when novel
324 combinations of species and environmental conditions may emerge[33,66].

325 **Conclusions**

326 In conclusion, we show that, although biodiversity experiments deliberately include plant communities
327 that may not currently occur under real-world conditions, the biodiversity-functioning relationship is
328 generally robust to the exclusion of these communities. Sensitivity analyses suggest that, where
329 biodiversity-functioning relationships did become weaker and non-significant, this change was
330 primarily driven by the truncated species-richness gradient in the realistic subset of experimental plots.
331 This indicates that experiments do not overestimate possible biodiversity-functioning relationships, but
332 rather that some real-world biodiversity gradients may not currently span the gradient in which
333 biodiversity loss has its strongest impact. Nevertheless, it is conceivable that future changes to
334 biodiversity may occur over this low to very low range.

335 Although we do not provide direct evidence for strong biodiversity-functioning relationships in
336 real-world communities, our results complement previous reports of significant biodiversity-
337 functioning relationships in the real world[4,36,42,57,58,63] by showing that constraining experimental datasets
338 to contain only real-world comparable plant communities does not change the core conclusions of
339 biodiversity-functioning research. However, to advance this field, we must acknowledge both the

strengths and limitations of biodiversity experiments. Specifically, our improved understanding should be used to develop a new generation of experiments, e.g. that focus on more realistic patterns of community change. At the same time, we must maintain and further examine the valuable resource of long-term biodiversity experiments, e.g. by re-analyzing existing experimental data to simulate a range of possible biodiversity-change scenarios. By moving beyond critiques of experimental design and placing experimental biodiversity-functioning research in the context of natural communities, we advance the current debate from verbal arguments to a quantitative investigation, thus increasing the robustness and applicability of biodiversity-functioning research.

Methods

1. Overview and data origin

We chose two of the largest and longest-running grassland biodiversity experiments in the world for our comparison. The Jena Experiment[6,32] was chosen as a Central-European example of a long-term, intensively studied biodiversity experiment[32,67]. In the Jena "main" experiment, combinations of 1, 2, 4, 8, 16 and 60 species from a pool of 60 Arrhenatherion grassland species[68] were sown in 82 originally 20 m × 20 m plots on a former agricultural field in 2002. This species richness gradient was crossed with a gradient of functional group richness (1 to 4 functional groups; small herbs, tall herbs, grasses, legumes), where species were randomly chosen from the respective functional groups (Roscher *et al.* 2004). Jena Experiment plots are maintained by weeding (two or three times per year). All plots are mown twice per year and mown biomass is removed, a common management practice of meadows in the region, and do not receive any fertilizers. The Jena Experiment includes two invasion sub-experiments, which are nested within the main experiment plots as subplots; one set of these Jena "invasion" plots was not weeded after initial sowing and studied regularly until 2009, another set was

362 weeded initially, but weeding halted in 2010[32]; here, we use the former for 2003–2009 and the latter for
363 2010–2015. Jena mown "succession" plots were not initially sown and are excluded from all
364 management except for the mowing. These plots represent intermediate successional stages between
365 the biodiversity experiment and the real-world systems, so they were included in the multivariate
366 analysis of community-property overlap (Fig. 1). However, given that they are influenced by initial
367 sowing, and that vegetation surveys were performed using different methodology (see below), they
368 were not considered real-world counterparts when constraining the Jena Experiment to realistic plots
369 (see below).

370 As a real-world counterpart to the Jena Experiment, we chose the grassland plots of the
371 Biodiversity Exploratories project (hereafter: "German real world"). This large-scale, long-term
372 research project was established in 2006 to assess the effects of land-use intensity on biodiversity and
373 ecosystem functioning in three regions of Germany[41]. The 150 grassland plots measure 50 m × 50 m
374 and were selected to cover a wide and representative range of land-use intensities, here composed of
375 varying levels of mowing frequency, grazing intensity and fertilization[69]. Species richness in
376 Exploratories grasslands ranges from nine to 70 species, within a 4 m × 4 m subplot, across all years
377 used in our study (see Supporting Fig. S8 for details on land-use intensity in the Biodiversity
378 Exploratories plots and its impact on the comparability of experimental and real-world communities).
379 Exploratories data were augmented by the inclusion of data from 14 grasslands in the Saale river valley
380 near the Jena Experiment (unpublished data; hereafter: "Jena real-world"). These grasslands are usually
381 mown twice per year; most are unfertilized and some are moderately fertilized.

382 The Cedar Creek biodiversity experiment e120 (hereafter: "BioDIV";[5,26,38,70]) was selected as a
383 North-American example of a long-term biodiversity experiment, while a suite of other naturally-
384 assembled grasslands at Cedar Creek served as nearby real-world communities. BioDIV was

established in 1994, when 1, 2, 4, 8 or 16 species were randomly drawn from an 18-species pool and sown across 168 13 m × 13 m plots at the Cedar Creek Ecosystem Science Reserve in Minnesota, USA.

Several datasets of local experiments and observation plots served as local real-world comparison for BioDIV. Experiments e001 (hereafter: "Fertilization 1") and e002 (hereafter: "Fertilization 2") were set up in 1982 to study the long-term effects of fertilization with nitrogen and other nutrients, ranging from low rates of nutrient inputs that are similar to atmospheric N deposition rates to high rates of fertilization similar to that used in agriculture. They consist of 324 plots located across three successional grassland fields (324 plots = 2 fertilization experiments × 3 old fields × 9 fertilization treatments × 6 replicates) that differ in their age since abandonment from agriculture and 45 plots in one never-plowed oak savannah in Fertilization 1 (45 plots = 9 nutrient treatments × 5 replicates)[43]. Plot sizes were 4 m × 4 m in the younger fields and 2 m × 4 m in the oak savannah. In contrast to Fertilization 1, Fertilization 2 plots were agriculturally disked before receiving nutrient addition treatments. Plot-level species richness in the two fertilization studies ranged from one to 28 species across all years used in our study. Established in 1983 and 1989, the Cedar Creek project e014 (hereafter "Old field succession chronosequence") offers vegetation data from four to six observational transects in each of 23 different fields repeated seven times between 1983 and 2011 to study succession after agricultural abandonment[45]. Cedar Creek project e093 (hereafter: "Oak savannah"), established in 1991, offers data from 30 2 m × 2 m prairie opening plots of natural vegetation[71,72]. This combination of Cedar Creek datasets was chosen to represent a variety of real-world plant communities that were comparable to the BioDIV experiment. Note that while Central European grasslands depend on anthropogenic management (mowing, grazing) to prevent succession to forest, the US prairies are naturally fire-disturbed, hence the selection of agricultural plots as the German real-world grassland. Please note that while all above-described datasets were used to illustrate multivariate overlap in plant

community properties (**Fig. 1a,b,e,f**), only a subset was used to constrain the biodiversity experiment data to realistic plots as different vegetation-survey techniques in the old field succession chronosequence and the oak savannah datasets (transects and subplots) made these data relatively incomparable (**Fig. 1c,d,g,h**; see below). For an overview of the datasets used in this study and online resources to obtain the original data, see Table S1 in Supporting Information.

2. Plant-community properties

Vascular plant cover and biomass

In the Jena Experiment, vegetation surveys were performed annually in the second half of May on a 3 m × 3 m subplot of each plot and species-specific cover data was collected. Note that, in the Jena "main" plots, only target species (vascular plants originally sown in the respective plots) were recorded. Vegetation surveys of the invasion and succession plots were performed annually in 2 m × 2.25 m subplots (2003-2009) or 3 m x 3 m subplots (2010-2015), assessing all present species. We used Jena vegetation data from 2003–2015 (succession data only from 2003–2009). In the Biodiversity Exploratories (German real-world plots), species-specific vascular plant cover was estimated annually on a 4 m × 4 m subplot of each plot between Mid-May and Mid-June. Here, we used all data from 2008-2015. Data from the 3 m × 3 m vegetation surveys of Jena real-world plots was available for May 2011.

To test if the different vegetation survey areas in Jena and the Biodiversity Exploratories might bias the relative abundance of vascular plant species and thus the calculation of abundance-weighted community properties, a separate survey of 27 Biodiversity Exploratories plots, which covered a strong land-use intensity gradient, was performed by sampling species-specific cover in a series of nested 4 m × 4 m (16 m², comparable to Exploratories vegetation survey area), 3 m × 3 m (9 m², comparable to

Jena Experiment and Jena real world) and 2 m × 2 m (4 m², similar to Jena invasion and succession) subplots. As cover estimates did not show any sign of systematic variation (Supporting Information, Fig. S9), we concluded that the different survey areas were unlikely to bias our analysis for the relative-abundance weighted community properties. We also compared species richness for the 27 16 m² and 9 m² subplots using a paired t-test. This showed a significantly lower species richness in the smaller subplots. On average, the 9 m² subplots had only 89 % of the species richness of the 16 m² subplots. Down-scaling species-richness related community properties based on such a coarse relationship established for only a subset of plots in only one year seemed inappropriate. However, data show that our results should be robust to differing vegetation survey areas of the datasets included in our study as species richness and most other taxonomic diversity community properties (except for D2 and SEve) were removed from the multidimensional comparison (PCA approach) based on the assessment of variance inflation factors (see below).

For BioDIV, a combination of species-specific cover data (1996–2000) and species-specific aboveground peak biomass (2001–2015) data was used to calculate plant community relative abundance. Previous analyses have shown that this difference in methodology does not affect the conclusions of analyses investigating species-richness effects on biomass[73]. Cover estimates for BioDIV were obtained by averaging the estimates from four permanently-marked subplots (each 0.5 m × 1 m) within each plot. Species-specific biomass in BioDIV was obtained by annually clipping 0.1 m × 6 m strips on each plot, drying and sorting the resulting biomass to species.

For Fertilization 1 and Fertilization 2, species-specific plant aboveground biomass data was collected annually at peak biomass by clipping a 0.1 m × 3 m strip of vegetation per plot, sorting and drying it. Years 1982–2004 were used for Fertilization 1 and 1982–1991 for Fertilization 2 as these years maintained the original, balanced treatment design, which was later changed to add further

453 treatments. For the old field succession chronosequence plots, species-specific cover values were used
454 for seven years between 1983 and 2011. Each of the 23 fields had four transects (except for two fields
455 with six transects) of 25 subplots each. For comparability to the other datasets, the 25 transect subplots
456 of 0.5 m × 1 m in each transect were treated as one plot by averaging species-specific cover values
457 across the subplots within transects resulting in four (or six) plots for each of the 23 fields (96 plots=21
458 fields × 4 plots + 2 fields × 6 plots). For the oak savannah dataset, only plant species cover from 1991
459 was used; later years were excluded because they were affected by a seed addition treatment. Species-
460 specific cover was averaged across the 16 0.5 m × 0.5 m subplots per plot.
461 For comparative analyses, different years were chosen for these different datasets due to
462 varying availability of measurements and to ensure a consistently-balanced design of the experimental
463 treatments in cases where additional treatments were added at a later stage. The transects in the old
464 field succession chronosequence are likely to inflate certain community properties because their
465 subplots span out further across the respective sites than a square plot of the same area would.
466 Similarly, the averaging across subplots in the oak savannah dataset might influence the direct
467 comparability to the biodiversity experiment data. As such, data from the old field succession
468 chronosequence and the oak savannah dataset are shown in **Fig. 1e** to put the BioDIV data into
469 perspective by adding different kinds of real-world data. However, when it came to constraining
470 biodiversity experiment data with the real-world data (**Fig. 1g**), we took a conservative approach and
471 included only those real-world datasets that were most comparable in terms of survey methodology
472 (Fertilization 1 and 2; hereafter: Combined US real world). Similarly, for the Jena Experiment real-
473 world counterparts, we considered only the German real world and Jena real-world plots as purely non-
474 biodiversity experiment plots in **Fig. 1c** (hereafter: Combined German real world).
475 To enable direct comparisons of plant communities, species-specific cover and biomass values

for all projects were transformed to relative abundance where the single abundance values within each community sum to 100. In order to do this, all Jena Experiment cover values (originally estimated on a decimal scale[74]) were first transformed to percent cover values[75]. Where vegetation covered more or less than 100% of the vegetation survey area (29 % of all communities in the German dataset had total cover below 100 %), it was scaled to 100% for the calculation of relative abundance and, subsequently, community properties. Some communities had a low overall cover, indicating bare ground. Specifically, although communities with a high percentage of bare ground were present in both experiments and the real-world, they were more common in the German biodiversity experiment than in its real-world counterparts. An equivalent assessment in the US datasets was not possible as relative abundance was here based on biomass rather than cover data for most communities - see above). Removing high-bare ground communities, where possible, might have led to an arbitrary, artificial convergence of plant-community properties from biodiversity experiments and real-world communities that would have weakened the direct comparison between those plant communities, a central aim of this study. Consequently, all communities were retained in the analysis.

Species synonyms and phylogeny

As we used plant species cover, biomass, and trait data from multiple sources based on research across decades and different geographic regions, there was considerable variation in the classification and nomenclature of species. Additionally, since the TRY database[76] was queried for plant traits and we also used a phylogenetic backbone tree (see below), the various datasets contained species names that might not all currently have the status of "accepted" names, challenging the linkage of the different datasets. This issue was dealt with by creating "code" data frames that linked all original spellings, outdated and synonym names which appeared in original data files to the respective accepted species names obtained

498 using The Plant List via function "TPL" in R package "Taxonstand"[77].

499 To calculate phylogenetic diversity metrics and to use phylogenetic relatedness to assist the
500 imputation of missing trait data, a phylogenetic tree of all plant species was created and included in our
501 study. We adopted the nomenclatural criteria in The Plant List v. 1.1[78] for the species in our dataset, and
502 pruned the updated vascular plant megaphylogeny by Qian & Jin[79] to include only the species in our
503 study (n = 664). We used the software SUNPLIN[80] to add the species lacking from the megaphylogeny
504 (n=132 or 19.9% of all species in our study) at random within the crown nodes of the corresponding
505 monophyletic genera. In a few cases where the genera of the missing species were polyphyletic
506 (*Potentilla, Medicago, Solidago, Galium*) or paraphyletic (*Calamagrostis, Vicia*), we inserted the
507 species at random within the nodes representing the most recent common ancestors that unequivocally
508 contain them (see[81]). We repeated this procedure iteratively to obtain 50 phylogenetic trees (see
509 Supporting Information, Fig. S10 for one example tree and the distribution of randomly inserted
510 species). When using the phylogenetic trees in the subsequent data analysis (calculation of
511 phylogenetic diversity metrics and plant trait imputation), all 50 trees were used and results were
512 averaged.

513 *Functional trait data*

514 In order to calculate community weighted mean trait values for all plant communities, functional trait
515 data from the TRY database (see Supporting Information, Table S15) were complemented with in-situ
516 collected trait data from Cedar Creek and not published in TRY. Plant species specific functional trait
517 values were calculated separately for the German and US species subsets.

518 Trait data for leaf area (mm^2), leaf dry mass (mg), leaf dry matter content (LDMC, g/g), leaf
519 nitrogen concentration (leaf N, mg/g), leaf phosphorus concentration (leaf P, mg/g), plant height (m),

specific leaf area (SLA, mm²/mg) and seed mass (dry mass in mg) were assembled[82]. These traits were selected as they are important for ecosystem functioning[46,47] and data for them was available. For the details of processing TRY and other trait data to generate species-level values, see Supporting Methods.

To fill gaps in trait data, trait values from same-genus species with available trait information were inferred. Subsequently, the "phylopars" function in the R package "Rphylopars"[83] was employed to impute missing data based on available information on other traits and the phylogenetic tree[84]. Before imputation, all trait data was natural-log transformed. To account for phylogenetic uncertainty (see above), trait data for all 50 phylogenetic trees was imputed and averaged. Subsequently, the plant species and their trait values were visualized in a PCA for each region (Supporting Information, Fig. S11) to check for strong outliers and check the outlier-species' ability to score extreme values. For details on the importance of species without original trait data (before genus inference and imputation) and for the number of species with identical trait information after inference and imputation, see Supporting Information Table S16.

Calculation of plant-community properties

Before calculating plant-community properties, tree species, occurring as seedlings, were removed from all datasets. This was because of their strong impact on the calculated CWM's and functional metrics, due to strong differences in trait expression between sapling (observed in the grasslands) and adult trees (studied for functional traits), and the fact that most grasslands in these climates, including the experiments, are grazed, mown or burned regularly, thus preventing tree invasion. Plant-community properties were calculated for each plot-year combination so that the temporal development (succession) of plots was accounted for in our analysis. As taxonomic diversity indices, we calculated species richness (S), Shannon's diversity (H), Simpson's diversity index (D1), and inverse Simpson's

diversity index (D2) (calculated as D1=1-D and D2=1/D, where D is the sum over all pi^2 and pi are the relative abundances of all species i) with functions "specnumber" and "diversity" in R package "vegan"[85] and Simpson's evenness (SEve, by dividing D2 by S)[86–89]. As phylogenetic diversity indices, we used Faith's phylogenetic diversity (PD), mean pairwise distance (MPD), and mean nearest taxon distance (MNTD)[90] with functions "pd", "mpd" and "mntd" in R package "picante"[91], where MPD and MNTD were calculated with abundance-weighting. All three phylogenetic diversity properties were calculated for each of the 50 phylogenetic trees and averaged to account for phylogenetic uncertainty (see above). For the calculation of the functional diversity indices functional richness (FRic), functional evenness (FEve), functional divergence (FDiv), functional dispersion (FDis), Rao's quadratic entropy (RaoQ)[92–94] and community weighted mean traits (CWM's) the function "dbFD" in the R package "FD"[93,95] was used with correction method "cailliez". As function "dbFD" relies on the computation of a Gower dissimilarity matrix where zero-dissimilarity values between two species (identical trait values) are not allowed, we slightly altered the trait values of a small number of species by deliberately increasing all trait values by 0.001 to 0.002% for the function to run. For each of the respective species pairs, only the species with the lower overall cover (throughout the regional dataset) received this alteration (Supporting Information, Table S17). For all but FRic, the abundance-weighted versions of these indices were computed. Communities comprising less than three species were assigned a value of zero for FRic, FEve, FDiv, PD, MPD and MNTD, as their computation is not possible for such communities.

3. Multivariate analysis of experiment and real-world intersection

Multivariate comparison

All analyses were carried out in R version 3.4.2[96]. Here, a multivariate PCA approach was employed,

564 based on numerous plant-community properties to assess the distribution, similarities and differences
565 between plant communities of biodiversity experiments and real-world systems. Our approach is based
566 on the relative distribution of plant communities in multidimensional, multivariate space. As this
567 distribution is highly-dependent on the community properties entering the PCAs and the information
568 they carry, we took care to avoid multicollinearity[97] among these community properties, to not over-
569 amplify information shared by several community properties. To quantitatively assess which variables
570 carried redundant information, we tested for multicollinearity of community properties by calculating
571 variance inflation factors (hereafter: vif; R function "corvif" provided by[98]). In the German and US
572 dataset, we sequentially removed the variables with the highest variance inflation factor until all vif
573 values were <3. Only the last of the eight variables to remove differed between the German and US
574 datasets, so for comparability between regional datasets, we removed all nine variables from both
575 datasets (see Supporting Information, Table S18 and S19). Specifically, H, FDis, S, leaf area, D1, PD,
576 MPD, RaoQ and FDiv were removed (in order of sequential removal) and only the following 12
577 community properties were employed in the PCA's: D2, SEve, FRic, FEve, SLA, leaf dry mass, leaf N,
578 leaf P, seed mass, height, LDMC, and MNTD (**Fig. 1b and f**; see Supporting Information Tables S20
579 and S21 for variance explained by all PCA axes and scores of the 12 community properties for the first
580 two axes, respectively). This vif-justified removal of community properties that were highly correlated
581 with S also helps with the issue of differences in species richness being caused by differing vegetation-
582 survey areas in the German real-world and Jena Experiment communities (see above). To test what
583 impact the selection of community properties entering the PCA had on our results, we re-ran our
584 analysis using various subsets of community properties or all of them (see below, Supporting
585 information on sensitivity analyses I, and Tables S2, S3, and S4). Separate community property PCA's
586 were computed for the German and USA data subsets using the "rda" function in R package "vegan"

(with variables scaled to avoid bias due to different range-size of properties) and the data was visualized in biplots with 95% confidence ellipses (**Fig. 1a and e**, see Supporting Information Table S22 for full dataset entering the PCA's).

Intersection-calculation methods

The intersection between experimental and real-world plots was calculated using three different methods of differing complexity, all based on the community-property PCA's presented in **Fig. 1a and e**. Intersections were calculated between two groups of data per geographic region: a) all experimental communities across all years and b) a subset of the most comparable and data-rich real-world datasets (combined real-world datasets). As described above, for Jena, the related combined real-world communities used in this intersection analysis were only the German real-world communities (Biodiversity Exploratories) and the Jena real-world communities. For BioDIV, only Fertilization 1 and Fertilization 2 plots were used as the combined real-world counterparts when calculating the intersections. First, the first two PCA axes were used to assess the two-dimensional intersection of 95% confidence ellipses for experimental and real-world data using the functions "ellipse" and "point.in.polygon" in R packages "car"[99] and "sp"[100,101], respectively (Supporting Information, Fig. S1). Second, the first three PCA axes were employed to compute the intersection of three-dimensional convex hull volumes using functions "convhulln" and "tsearchn" in R package "geometry"[102] (**Fig. 1c and g** show 2-dimensional representation of 3-dimensional convex hull volume). Third, using the first three PCA axes, three-dimensional hypervolumes were computed using the "hypervolume" package in R[103]. The intersection hypervolume of the experimental and real-world hypervolumes was then calculated and function "hypervolume_inclusion_test" was used to assess which communities fall in the intersection hypervolume (Supporting Information, Fig. S1). For the subsequent analysis of

609 diversity-functioning (hereafter: BEF) relationships, experimental plots were defined as realistic if their
610 plant communities fell inside the intersection in at least one of the years present in the dataset. Higher
611 thresholds (e.g., 90 % of the years inside the intersection) may be inappropriate given that the early
612 years of the experiment see the establishment of sown communities, and would have rendered too few
613 Jena Experiment plots realistic to adequately assess biodiversity-functioning relationships in
614 constrained datasets (Supporting Information on sensitivity analyses I and Tables S3 and S4). As such,
615 the inclusion criterion used resulted in the selection of the most realistic experimental plots, while also
616 providing a sufficient number of realistic plots to compare biodiversity-functioning relationships in
617 constrained and unconstrained datasets. Given this threshold, each plot in the experiments was either
618 defined as realistic (the plot's plant community was within the intersection in at least one year) or
619 unrealistic. Calculating the intersection based on three different methods of different complexity
620 demonstrated that the selection of realistic communities was largely insensitive to the underlying
621 methodology (Supporting Information, Table S3 and S4). Therefore, we focus our analyses on using
622 three-dimensional convex-hull volumes, a method of intermediate complexity, and present results for
623 the other methods in the Supporting Information.

624 **4. Measurement of ecosystem-function variables**

625 A range of above- and belowground ecosystem process rates and state variables was selected as
626 ecosystem functions from the Jena Experiment and BioDIV in such a way that the functions of these
627 experiments were as comparable as possible. Only function data obtained between 2006 and 2015 (at
628 least 4 years after initiation of the experiments) was used because BEF relationships shortly after the
629 initial establishment of experiments are often unrepresentative of longer-term trends[26,104]. These
630 selection criteria resulted in the following functions: Plant aboveground biomass (biomass),

631 aboveground plant biomass C:N ratio (plant C:N), soil organic carbon (C) and root biomass were
632 available for both experiments. As soil inorganic C should not play a role at BioDIV due to the sandy
633 soil, measurements of total C can be considered representative of organic C stocks here (see Supporting
634 Methods). Herbivory rate, soil microbial biomass C, phosphatase activity, and pollinator abundance
635 were only available for Jena. For details regarding the measurement of these ecosystem functions in the
636 Jena Experiment and BioDIV; please refer to the Supporting Methods section.

637 **5. Statistical analysis of unconstrained and constrained experimental BEF relationships**
638 In order to assess whether – and how much – BEF relationships change when excluding unrealistic
639 plots from the analysis, each relationship was first analyzed in the unconstrained dataset with all
640 experimental plots. Subsequently, biodiversity experiment datasets were constrained to only include
641 realistic plots and the models were re-run. For ecosystem function variables with multiple years of data,
642 values were averaged across years and simple linear models were fit that tested for the effect of realized
643 target species richness ($\log_2$, averaged per plot between 2006 and 2015) on the individual functions.
644 Where necessary, square-root or $\log_{10}$-transformation was applied to response variables to meet model
645 assumptions of normality and homoscedasticity of variances. For each of the resulting relationships,
646 slope estimates and their 95% confidence intervals (function "confint" in R) were calculated. Slopes
647 and confidence intervals of each pair of constrained and unconstrained relationships were compared to
648 decide if the slope or sign of the relationship had changed. If confidence intervals of unconstrained and
649 constrained slopes included each other's mean value, we concluded that they were not significantly
650 different. Additionally, a paired t-test, directly comparing the slope values estimated from
651 unconstrained and constrained data subsets (for the twelve BEF relationships in Fig. 2, n=12) was
652 performed.

653 **6. Sensitivity analyses**

654 Since our analysis involved many decisions on which variables to include and what exact analytical
655 pathway to follow, and these decisions might affect our results, several sensitivity analyses were
656 performed regarding different aspects of our analysis.

657 To test if different subsets of community properties entering the PCA affected our results, our
658 analysis was re-run for combinations of i) different subsets of community properties, i.e. a) the vif-
659 selected 12 community properties (presented in the main text), b) all available 21 community
660 properties, and c) four subsets excluding one class of community properties (taxonomic, phylogenetic,
661 functional diversity, or CWM functional traits, respectively) and ii) three methods to compute the
662 intersection between biodiversity experiment and real-world plots described above (Supporting
663 Information, Fig. S1 and S6). These community-property subsets were used to demonstrate how
664 strongly the results were influenced by each class of community properties. To keep the number of
665 sensitivity analyses manageable given the high number of possible combinations of community
666 properties and overlap calculation methods, only the vif-selected subset and the set containing all 21
667 community properties were tested with all three methods. Additionally, we conducted a series of
668 sensitivity analyses that assessed the impact of other methodological changes on the PCA-based
669 selection of realistic biodiversity experiment plots. They include: using more subsets of community
670 properties (sensitivity analysis A), including more principal components (axes) of the PCA to define
671 realistic plots based on higher-dimensional space (B), including all available real-world datasets (not
672 just the most methodologically comparable ones, C), using different inclusion criteria to define
673 experimental plots as realistic (D), using species-abundance based NMDS rather than community-
674 property based PCA's to assess intersections of different datasets (E, Supporting Information Fig. S5)

675 and including only those German real-world plots in the PCA's that resemble the Jena Experiment in
676 their land use (F). Details on the methodology and results of these sensitivity analyses are described in
677 the Supporting Information on sensitivity analyses I, Tables S2, S3, and S4, and Figures S1, S5, S6.
678 To test if shifts in significance of BEF relationships in **Fig. 2** simply resulted from the strong
679 reduction of error degrees of freedom associated with using data subsets, we performed a sensitivity
680 analysis in which we randomly selected the same proportion of plots as realistic as that in our PCA-
681 driven selection of realistic plots, 500 times for each relationship (Supporting Information on
682 sensitivity analyses II, Fig. S7a). In addition, we performed an alternative version of this sensitivity
683 analysis that restricted the random draws of Jena Experiment plots to only those with a species richness
684 falling within the truncated species-richness gradient of the realistic Jena plots (Fig. S7b).
685 To gain further insight into our findings at Jena, data from experimental plots which were
686 abandoned and allowed to undergo natural succession (Jena invasion plots) were more closely
687 analyzed. Over time, these migrated towards the multivariate community-property space occupied by
688 real-world communities (Supporting Information, Fig. S2).

Acknowledgements

We thank the establishers, maintainers, coordinators, technical and research staff, and data owners of all involved projects, the TRY initiative and data owners. Santiago Soliveres, Eric Allan for discussion, and Sven Thiel, Guangjuan Luo and Dan Bahauddin, and Florian Schneider for help with data extraction and handling, Robert Junker and Benjamin Blonder assisted with the calculation of multidimensional hypervolumes. This study was funded through Jena Experiment SP 7 (Swiss National Science Foundation 310030E-166017/1). Further support came from the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, funded by the German Research Foundation (FZT 118). The Jena Experiment was funded by the Deutsche Forschungsgemeinschaft (FOR 456 and FOR 1451) with additional support from Friedrich Schiller University Jena, Max Planck Institute for Biogeochemistry in Jena, and Swiss National Science Foundation. All Cedar Creek studies are funded by the US National Science Foundation's Long-Term Ecological Research (LTER) program (DEB-1234162). FI acknowledges funding from the LTER Network Communications Office (DEB-1545288). We are grateful for the constructive feedback and suggestions by three anonymous reviewers.

Data accessibility

We provide aggregated datasets with plant-community properties and ecosystem function data at first submission to enable editors and referees to run our main analyses. Currently, these datasets partly underlie project-specific embargo periods and need to be treated confidentially. All data will be a) uploaded to an online repository, b) submitted as supplemental files upon acceptance of the article or c) be made available within project databases after the respective project-defined embargo periods. Upon request by editors or referees, we are happy to provide all data at an earlier stage.

Code availability

We provide R-code for running the main analyses and creating Fig. 1 and Fig. 2 based on aggregated datasets at first submission. All R-code for data crunching and analyses will be a) uploaded to an online repository, b) submitted as supplemental files upon acceptance of the article or c) be made available within project databases after the respective project-defined embargo periods. Upon request by editors or referees, we are happy to provide all R-code at an earlier stage.

Author contributions

MJ, PM, MF and FvdP conceived and designed the study; all authors except for FvdP, RM-V, CP and AR contributed data; MJ developed the analytical framework and analyzed the data; RM-V constructed the phylogenetic hypothesis trees; MJ and PM wrote the manuscript; all authors contributed to the discussion of results and writing of the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Cardinale, B. J. *et al.* Biodiversity loss and its impact on humanity. *Nature* **486**, 59–67 (2012).
2. Tilman, D., Isbell, F. & Cowles, J. M. Biodiversity and ecosystem functioning. *Annu. Rev. Ecol. Evol. Syst.* **45**, 471–493 (2014).
3. Isbell, F. *et al.* Linking the influence and dependence of people on biodiversity across scales. *Nature* **546**, 65–72 (2017).
4. van der Plas, F. Biodiversity and ecosystem functioning in naturally assembled communities. *Biol. Rev.* **online-ear**, (2019).
5. Tilman, D. *et al.* Diversity and productivity in a long-term grassland experiment. *Science.* **294**,

843–845 (2001).

6. Roscher, C., Schumacher, J. & Baade, J. The role of biodiversity for element cycling and trophic interactions: an experimental approach in a grassland community. *Basic Appl. Ecol.* **121**, 107–121 (2004).

7. Hooper, D. U. *et al.* Effects of biodiversity on ecosystem functioning: a consensus of current knowledge. *Ecol. Monogr.* **75**, 3–35 (2005).

8. Cardinale, B. J. *et al.* The functional role of producer diversity in ecosystems. *Am. J. Bot.* **98**, 572–592 (2011).

9. O'Connor, M. I. *et al.* A general biodiversity–function relationship is mediated by trophic level. *Oikos* **126**, 18–31 (2017).

10. Huston, M. A. Hidden treatments in ecological experiments: re-evaluating the ecosystem function of biodiversity. *Oecologia* **110**, 449–460 (1997).

11. Grime, J. P. Benefits of plant diversity to ecosystems: Immediate, filter and founder effects. *J. Ecol.* **86**, 902–910 (1998).

12. Wardle, D. A. *et al.* Biodiversity and ecosystem function: an issue in ecology. *Bull. Ecol. Soc. Am.* **81**, 235–239 (2000).

13. Leps, J. What do the biodiversity experiments tell us about consequences of plant species loss in the real world? *Basic Appl. Ecol.* **5**, 529–534 (2004).

14. Srivastava, D. S. & Vellend, M. Biodiversity-ecosystem function research: is it relevant to conservation? *Annu. Rev. Ecol. Evol. Syst.* **36**, 267–294 (2005).

15. Duffy, J. E. Why biodiversity is important to the functioning of real-world ecosystems. *Front. Ecol. Environ.* **7**, 437–444 (2008).

16. Duffy, J. E. Biodiversity effects: trends and exceptions – a reply to Wardle and Jonsson. *Front. Ecol. Environ.* **8**, 11–12 (2010).

17. Wardle, D. A. & Jonsson, M. Biodiversity effects in real ecosystems – a response to Duffy. *Front. Ecol. Environ.* **8**, 10–11 (2010).

18. Wardle, D. A. Do experiments exploring plant diversity-ecosystem functioning relationships inform how biodiversity loss impacts natural ecosystems? *J. Veg. Sci.* **27**, 646–653 (2016).

19. Manning, P. *et al.* Transferring biodiversity-ecosystem function research to the management of 'real-world' ecosystems. in *Advances in Ecological Research 61* 323–356 (2019).

20. Schulze, E.-D. & Mooney, H. *Biodiversity and Ecosystem Functioning*. (Springer, 1993).

21. Naeem, S., Thompson, L. J., Lawler, S. P., Lawton, J. H. & Woodfin, R. M. Declining biodiversity can alter the performance of ecosystems. *Nature* **368**, 734–737 (1994).

22. Balvanera, P. *et al.* Quantifying the evidence for biodiversity effects on ecosystem functioning and services. *Ecol. Lett.* **9**, 1146–56 (2006).

23. Hines, J. *et al.* Mapping change in biodiversity and ecosystem function research: food webs foster integration of experiments and science policy. *Adv. Ecol. Res.* **61**, 297–322 (2019).

24. Tilman, D., Wedin, D. & Knops, J. Productivity and sustainability influenced by biodiversity in grassland ecosystems. *Nature* **379**, 718–720 (1996).

25. Loreau, M. *et al.* Biodiversity and Ecosystem Functioning: Current Knowledge and Future Challenges. *Science.* **294**, 804–808 (2001).

26. Reich, P. B. *et al.* Impacts of biodiversity loss escalate through time as redundancy fades. *Science.* **336**, 589–592 (2012).

27. Wilsey, B. J. & Potvin, C. Biodiversity and ecosystem functioning: Importance of species evenness in an old field. *Ecology* **81**, 887–892 (2000).

28. Wilsey, B. J. & Polley, H. W. Realistically Low Species Evenness Does Not Alter Grassland Species-Richness–Productivity Relationships. *Ecology* **85**, 2693–2700 (2004).

29. Hillebrand, H., Bennett, D. & Cadotte, M. Consequences of dominance: A Review of evenness effects on local and regional ecosystem processes. *Ecology* **89**, 1510–1520 (2008).

30. Schmitz, M. *et al.* Consistent Effects of Biodiversity on Ecosystem Functioning Under Varying Density and Evenness. *Folia Geobot.* **48**, 335–353 (2013).

31. Finn, J. A. *et al.* Ecosystem function enhanced by combining four functional types of plant species in intensively managed grassland mixtures: A 3-year continental-scale field experiment. *J. Appl. Ecol.* **50**, 365–375 (2013).

32. Weisser, W. W. *et al.* Biodiversity effects on ecosystem functioning in a 15-year grassland experiment: patterns, mechanisms, and open questions. *Basic Appl. Ecol.* **23**, 1–73 (2017).

33. Schmid, B. & Hector, A. The value of biodiversity experiments. *Basic Appl. Ecol.* **5**, 535–542 (2004).

34. Eisenhauer, N. *et al.* Biodiversity-ecosystem function experiments reveal the mechanisms underlying the consequences of biodiversity change in real world ecosystems. *J. Veg. Sci.* **27**, 1061–1070 (2016).

35. Isbell, F. *et al.* Nutrient enrichment, biodiversity loss, and consequent declines in ecosystem productivity. *Proc. Natl. Acad. Sci.* **110**, 11911–11916 (2013).

795 36. Duffy, J. E., Godwin, C. M. & Cardinale, B. J. Biodiversity effects in the wild are common and
796 as strong as key drivers of productivity. *Nature* **549**, 261–264 (2017).

797 37. Buchmann, T. *et al.* Connecting experimental biodiversity research to real-world grasslands.
798 *Perspect. Plant Ecol. Evol. Syst.* **33**, 78–88 (2018).

799 38. Tilman, D. *et al.* The influence of functional diversity and composition on ecosystem processes.
800 *Science.* **277**, 1300–1302 (1997).

801 39. Tilman, D., Reich, P. B. & Isbell, F. Biodiversity impacts ecosystem productivity as much as
802 resources, disturbance, or herbivory. *Proc. Natl. Acad. Sci.* **109**, 10394–10397 (2012).

803 40. Isbell, F. *et al.* Biodiversity increases the resistance of ecosystem productivity to climate
804 extremes. *Nature* **526**, 574–577 (2015).

805 41. Fischer, M. *et al.* Implementing large-scale and long-term functional biodiversity research: The
806 Biodiversity Exploratories. *Basic Appl. Ecol.* **11**, 473–485 (2010).

807 42. Soliveres, S. *et al.* Biodiversity at multiple trophic levels is needed for ecosystem
808 multifunctionality. *Nature* **536**, 456–459 (2016).

809 43. Tilman, D. Secondary succession and the pattern of plant dominance along experimental
810 nitrogen gradients. *Ecol. Monogr.* **57**, 189–214 (1987).

811 44. Clark, C. M. & Tilman, D. Loss of plant species after chronic low-level nitrogen deposition to
812 prairie grasslands. *Nature* **451**, 712–715 (2008).

813 45. Inouye, R. *et al.* Old-field succession on a Minnesota sand plain. *Ecology* **68**, 12–26 (1987).

814 46. Díaz, S. *et al.* The global spectrum of plant form and function. *Nature* **529**, 1–17 (2015).

815 47. Craven, D. *et al.* Multiple facets of biodiversity drive the diversity–stability relationship. *Nat.*
816 *Ecol. Evol.* **2**, 1579–1587 (2018).

817 48. Nakamura, G., Gonçalves, L. O. & Duarte, L. da S. Revisiting the dimensionality of biological
818 diversity. *Ecography (Cop.).* 1–10 (2019). doi:10.1111/ecog.04574

819 49. Stevens, R. D. & Tello, J. S. On the measurement of dimensionality of biodiversity. *Glob. Ecol.*
820 *Biogeogr.* **23**, 1115–1125 (2014).

821 50. Manning, P. *et al.* Simple measures of climate, soil properties and plant traits predict national-
822 scale grassland soil carbon stocks. *J. Appl. Ecol.* **52**, 1188–1196 (2015).

823 51. Loreau, M. & Hector, A. Partitioning selection and complementarity in biodiversity experiments.
824 *Nature* **412**, 72–76 (2001).

825 52. Allan, E. *et al.* Land use intensification alters ecosystem multifunctionality via loss of

biodiversity and changes to functional composition. *Ecol. Lett.* **18**, 834–843 (2015).

53. Le Bagousse-Pinguet, Y. *et al.* Phylogenetic, functional, and taxonomic richness have both positive and negative effects on ecosystem multifunctionality. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 8419–8424 (2019).

54. Venail, P. *et al.* Species richness, but not phylogenetic diversity, influences community biomass production and temporal stability in a re-examination of 16 grassland biodiversity studies. *Funct. Ecol.* **29**, 615–626 (2015).

55. Hillebrand, H. & Matthiessen, B. Biodiversity in a complex world: consolidation and progress in functional biodiversity research. *Ecol. Lett.* **12**, 1405–1419 (2009).

56. Grace, J. B. *et al.* Integrative modelling reveals mechanisms linking productivity and plant species richness. *Nature* **529**, 390–393 (2016).

57. Liang, J. *et al.* Positive biodiversity-productivity relationship predominant in global forests. *Science.* **354**, aaf8957 (2016).

58. Oehri, J., Schmid, B., Schaepman-Strub, G. & Niklaus, P. A. Biodiversity promotes primary productivity and growing season lengthening at the landscape scale. *Proc. Natl. Acad. Sci.* **114**, 10160–10165 (2017).

59. Díaz, S. *et al.* Incorporating plant functional diversity effects in ecosystem service assessments. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 20684–20689 (2007).

60. Lavorel, S. *et al.* Using plant functional traits to understand the landscape distribution of multiple ecosystem services. *J. Ecol.* **99**, 135–147 (2011).

61. Schmid, B. The species richness-productivity controversy. *Trends Ecol. Evol.* **17**, 113–114 (2002).

62. Loreau, M. Biodiversity and ecosystem functioning: A mechanistic model. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 5632–5636 (1998).

63. Maestre, F. T. *et al.* Plant species richness and ecosystem multifunctionality in global drylands. *Science.* **335**, 214–218 (2012).

64. van der Plas, F. *et al.* Jack-of-all-trades effects drive biodiversity-ecosystem multifunctionality relationships in European forests. *Nat. Commun.* **7**, 11109 (2016).

65. Socher, S. A. *et al.* Direct and productivity-mediated indirect effects of fertilization, mowing and grazing on grassland species richness. *J. Ecol.* **100**, 1391–1399 (2012).

66. Hobbs, R. J., Higgs, E. & Harris, J. a. Novel ecosystems: implications for conservation and restoration. *Trends Ecol. Evol.* **24**, 599–605 (2009).

67. Roscher, C. *et al.* Convergent high diversity in naturally colonized experimental grasslands is not related to increased productivity. *Perspect. Plant Ecol. Evol. Syst.* **20**, 32–45 (2016).

68. Ellenberg, H. & Leuschner, C. *Vegetation Mitteleuropas mit den Alpen: in ökologischer, dynamischer und historischer Sicht*. (UTB, 2010).

69. Blüthgen, N. *et al.* A quantitative index of land-use intensity in grasslands: Integrating mowing, grazing and fertilization. *Basic Appl. Ecol.* **13**, 207–220 (2012).

70. Tilman, D., Reich, P. B. & Knops, J. M. H. Biodiversity and ecosystem stability in a decade-long grassland experiment. *Nature* **441**, 629–632 (2006).

71. Tilman, D. Community Invasibility, Recruitment Limitation, and Grassland Biodiversity. *Ecology* **78**, 81–92 (1997).

72. Catford, J. A. *et al.* Traits linked with species invasiveness and community invasibility vary with time, stage and indicator of invasion in a long-term grassland experiment. *Ecol. Lett.* **22**, 593–604 (2019).

73. Fargione, J. *et al.* From selection to complementarity: shifts in the causes of biodiversity-productivity relationships in a long-term biodiversity experiment. *Proc. R. Soc. B Biol. Sci.* **274**, 871–876 (2007).

74. Londo, G. The decimal scale for releves of permanent quadrats. *Vegetatio* **33**, 61–64 (1976).

75. Roscher, C. *et al.* What happens to the sown species if a biodiversity experiment is not weeded? *Basic Appl. Ecol.* **14**, 187–198 (2013).

76. Kattge, J. *et al.* TRY - a global database of plant traits. *Glob. Chang. Biol.* **17**, 2905–2935 (2011).

77. Cayuela, L., Stein, A. & Oksanen, J. Taxonstand: Taxonomic Standardization of Plant Species Names. *R Packag. version 2.1* (2017).

78. The Plant List, Version 1.1. (2013). Available at: http://www.theplantlist.org/.

79. Qian, H. & Jin, Y. An updated megaphylogeny of plants, a tool for generating plant phylogenies and an analysis of phylogenetic community structure. *J. Plant Ecol.* **9**, 233–239 (2016).

80. Martins, W. S., Carmo, W. C., Longo, H. J., Rosa, T. C. & Rangel, T. F. SUNPLIN: Simulation with Uncertainty for Phylogenetic Investigations. *BMC Bioinformatics* **14**, (2013).

81. Rangel, T. F. *et al.* Phylogenetic uncertainty revisited: Implications for ecological analyses. *Evolution (N. Y).* **69**, 1301–1312 (2015).

82. Cornelissen, J. H. C. *et al.* A handbook of protocols for standardised and easy measurement of plant functional traits worldwide. *Aust. J. Bot.* **51**, 335–380 (2003).

83. Goolsby, E. W., Bruggeman, J. & Ane, C. Rphylopars: Phylogenetic Comparative Tools for Missing Data and Within-Species Variation. (2016).

84. Penone, C. *et al.* Imputation of missing data in life-history trait datasets: Which approach performs the best? *Methods Ecol. Evol.* **5**, 1–10 (2014).

85. Oksanen, J. *et al.* Vegan: community ecology package. *R Packag. version 2.3-4* (2016).

86. Hill, M. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology* **54**, 427–432 (1973).

87. Smith, B. & Wilson, J. B. A Consumer's Guide to Evenness Indices. *Oikos* **76**, 70–82 (1996).

88. Magurran, A. *Measuring Biological Diversity*. (Blackwell, 2004).

89. Morris, E. K. *et al.* Choosing and using diversity indices: Insights for ecological applications from the German Biodiversity Exploratories. *Ecol. Evol.* **4**, 3514–3524 (2014).

90. Tucker, C. M. *et al.* A guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biol. Rev.* **92**, 698–715 (2017).

91. Kembel, S. W. *et al.* Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**, 1463–1464 (2010).

92. Villéger, S., Mason, N. W. H. & Mouillot, D. New multidimensional functional diversity indices for a multifaceted framework in functional ecology. *Ecology* **89**, 2290–2301 (2008).

93. Laliberte, E. & Legendre, P. A distance-based framework for measuring functional diversity from multiple traits A distance-based framework for measuring from multiple traits functional diversity. *Ecology* **91**, 299–305 (2010).

94. Mouchet, M. A., Villéger, S., Mason, N. W. H. & Mouillot, D. Functional diversity measures: an overview of their redundancy and their ability to discriminate community assembly rules. *Funct. Ecol.* **24**, 867–876 (2010).

95. Laliberté, E., Legendre, P. & Shipley., B. FD: measuring functional diversity from multiple traits, and other tools for functional ecology. (2014).

96. R Core Team. R: A Language and Environment for Statistical Computing. (2019). doi:10.1007/978-3-540-74686-7

97. Dormann, C. F. *et al.* Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography (Cop.).* **36**, 27–46 (2013).

98. Zuur, A. F., Ieno, E. N. & Elphick, C. S. A protocol for data exploration to avoid common statistical problems. *Methods Ecol. Evol.* **1**, 3–14 (2010).

99. Fox, J. & Weisberg, S. *An R Companion to Applied Regression*. (Thousand Oaks, 2011).

100. Pebesma, E. & Bivand, R. Classes and methods for spatial data. *R News* **5**, 9–13 (2005).

101. Bivand, R. S., Pebesma, E. & Gomez-Rubio, V. *Applied spatial data analysis with R*. (Springer, 2013).

102. Habel, K., Grasman, R., Gramacy, R. B., Stahel, A. & Sterratt, D. C. geometry: Mesh Generation and Surface Tesselation. *R package version 0.4.1* (2019).

103. Blonder, B. & Harris, D. hypervolume: High Dimensional Geometry and Set Operations Using Kernel Density Estimation, Support Vector Machines, and Convex Hulls. *R Packag. ver. 2.0.11* (2018).

104. Meyer, S. T. *et al.* Effects of biodiversity strengthen over time as ecosystem functioning declines at low and increases at high biodiversity. *Ecosphere* **7**, e01619 (2016).

Figures

Fig. 1 Experimental versus real-world communities. Upper row: German comparison (n=3,329 plot-year combinations). Lower row: US comparison (n=9,954 plot-year combinations). a-c and e-g: First two axes of a PCA on 12 plant-community properties (see panels b and f, variance-inflation factor-selected community weighted mean traits, functional diversity, phylogenetic diversity and taxonomic diversity metrics), where each dot represents a single plot in a single year. a and e: Distribution of the experimental (orange) and various real-world plots with 95% confidence ellipses (variables scaled for PCA) for each subset. b and f: PCA factor loadings for community properties (arrows proportionally increased to improve visibility - "const=25" in R vegan "biplot" function[85]; see Supporting Information, Table S21 and S22 for PCA factor loadings and the full dataset, respectively). c and g: Two-dimensional representation of three-dimensional convex hull volumes for experimental (orange) and combined real-world communities (German real world and Jena real-world plots for the German, Fertilization 1 and 2 plots for the US comparison, gray) and their intersection (shaded area). d and h: Number and proportion (strong versus light color) of biodiversity experiment plots in the intersection i.e. realistic plots, where each plot with at least one annual community in the intersection is defined as realistic. Number of years of vegetation data for each project: Jena Experiment (13), German real world (8), Jena real world (1), Jena invasion (13), Jena succession (7), BioDIV (19), Fertilization 1 (23) and 2 (10), Old field succession chronosequence (7), Oak savannah (1). Abbreviations of community properties: taxonomic diversity indices: inverse Simpson's diversity index (D2) and Simpson's evenness (SEve); phylogenetic diversity indices: mean nearest taxon distance (MNTD); functional diversity indices: functional richness (FRic), and functional evenness (FEve); CWM values of leaf nitrogen (Leaf_N) and phosphorus (Leaf_P), specific leaf area (SLA), leaf dry mass, leaf dry matter content (LDMC), seed mass and plant height. For definitions of these properties, please see Methods.

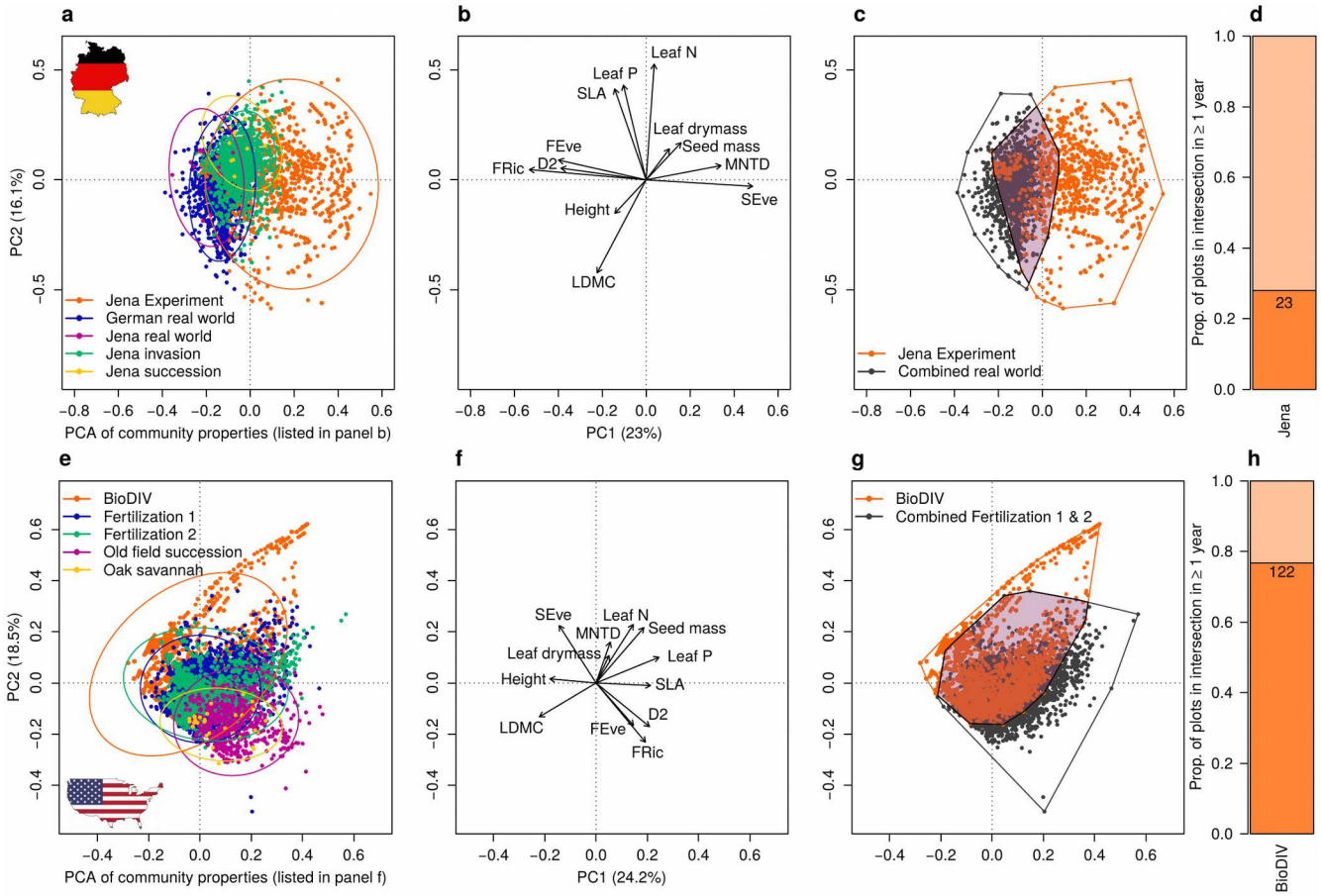
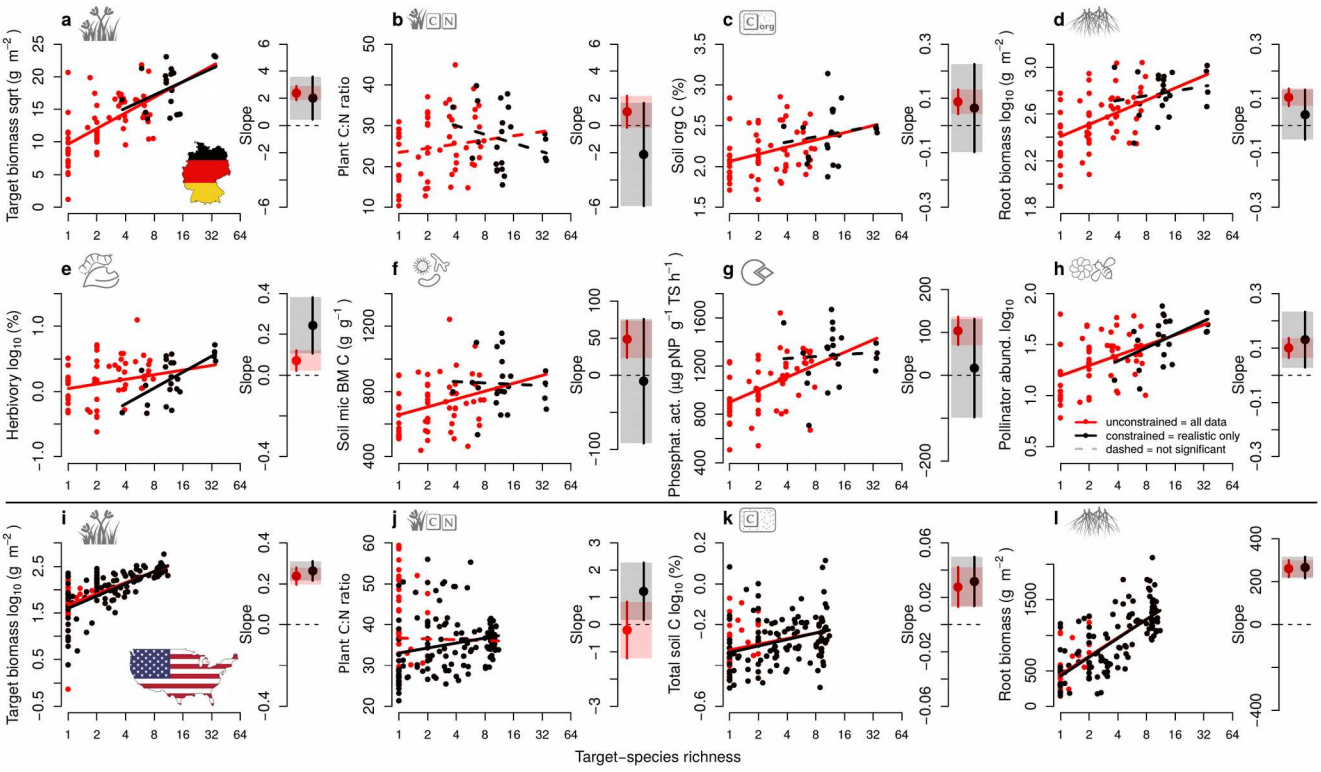

a
b
c
d
e
f
g
h
PC2 (16.1%)
PCA of community properties (listed in panel b)
Jena Experiment
German real world
Jena real world
Jena invasion
Jena succession
Leaf N
Leaf P
SLA
Leaf drymass
Seed mass
FEve
FRic
D2
MNTD
SEve
Height
LDMC
PC1 (23%)
Jena Experiment
Combined real world
Prop. of plots in intersection in ≥ 1 year
23
Jena
PC2 (18.5%)
PCA of community properties (listed in panel f)
BioDIV
Fertilization 1
Fertilization 2
Old field succession
Oak savannah
PC1 (24.2%)
BioDIV
Combined Fertilization 1 & 2
122
BioDIV

Fig. 2 Biodiversity-ecosystem functioning relationships. Relationship between realized target plant species richness (averaged per plot between 2006 and 2015, axis on log_2-scale) and various ecosystem functions in German (panels a-h, Jena Experiment) and US (panels i-l, BioDIV) biodiversity experiments containing all plots (unconstrained, all dots and red lines) and only realistic plots (constrained, black dots and lines). Insets show slope estimates with 95% confidence intervals (error bars and shaded areas) for all plots (unconstrained, red) and only realistic plots (constrained, black). For model parameters such as sample sizes, slope estimates, confidence intervals, p-values and adjusted R^2 values, see Supporting Information, Table S9. Dashed regression lines show non-significant relationships ($p>0.05$). Note that panels a-d and i-l show the same ecosystem functions for both experiments (in BioDIV, total soil C represents soil organic C, panel k). BM denotes biomass and C:N ratio means carbon to nitrogen ratio. Where indicated in the y-axis label, data were transformed to meet model assumptions. Response variables were averaged over all available years. Function symbols modified from originals by Hamish, Saeful Muslim, Alice Noir, Lluis Pareras, Creative Stall, Atif Arshad, Made and amantaka from the Noun Project.


a
b
c
d
e
f
g
h
i
j
k
l
Target biomass sqrt (g m^{-2})
Plant C:N ratio
Soil org C (%)
Root biomass log$_{10}$ (g m^{-2})
Herbivory log$_{10}$ (%)
Soil mic BM C (g g^{-1})
Phosphat. act. (µg pNP g^{-1} TS h^{-1})
Pollinator abund. log$_{10}$
Target biomass log$_{10}$ (g m^{-2})
Plant C:N ratio
Total soil C log$_{10}$ (%)
Root biomass (g m^{-2})
Slope
unconstrained = all data
constrained = realistic only
dashed = not significant
Target−species richness

969 **Supporting Information**

970 The following Supporting Information is available for this article online:

971 **Supporting Methods.**

972 **Table S1.** List of German and US datasets for vegetation and ecosystem function data.

973 **Supporting Information on sensitivity analyses I.**

974 **Table S2.** Community properties used in the different main and sensitivity subsets.

975 **Figure S1.** Alternative versions of Fig. 1 based on the alternative intersection scenarios.

976 **Table S3.** Jena plots included in the different overlap scenarios versus all experimental plots.

977 **Table S4.** BioDIV plots included in the different overlap scenarios versus all experimental plots.

978 **Figure S2.** Temporal movement of Jena invasion communities into the real-world realm.

979 **Figure S3.** Violin plots of community properties of German experimental and real-world plots.

980 **Table S5.** T-test results for differences between German experimental and real-world plots.

981 **Figure S4.** Violin plots of community properties of US experimental and real-world plots.

982 **Table S6.** T-test results for differences between US experimental and real-world plots.

983 **Figure S5.** NMDS biplots of species-abundance data for German and US dataset.

984 **Table S7.** T-test results for differences between realistic and unrealistic plots for the Jena Experiment.

985 **Table S8.** T-test results for differences between realistic and unrealistic plots for BioDIV.

986 **Table S9.** Model parameters for BEF relationships presented in Fig. 2.

987 **Figure S6.** Alternative versions of Fig. 2 based on the alternative intersection scenarios.

988 **Table S10.** Constraining-related change in functioning at maximum species richness.

989 **Supporting Information on sensitivity analyses II.**

990 **Figure S7:** Random selection sensitivity analysis for Fig. 2 relationships turning insignificant.

991 **Table S11.** Differences between range in function covered by unconstrained and constrained models in

Fig. 2.

Table S12. Drivers of ecosystem functioning in all vs. realistic communities of biodiversity experiments.

Table S13. Correlation coefficients for CWM's versus functional, phylogenetic metrics and evenness, German dataset.

Table S14. Correlation coefficients for CWM's versus functional, phylogenetic metrics and evenness, US dataset.

Figure S8. Alternative versions of Fig. 1a showing Exploratories land-use intensity gradients.

Figure S9. Cover versus vegetation survey size scaling sensitivity check for German real-world data (Biodiversity Exploratories).

Figure S10. Phylogenetic backbone tree (one example of the 50 replicates).

Table S15. TRY references for plant species trait data from two TRY requests (might have to be included in the main references depending on the TRY rules and journal policy).

Figure S11. PCA of plant species and their traits for German and US comparison.

Table S16. Percentage cover of species with trait information for the German and US datasets.

Table S17. Species with altered trait values to avoid Gower dissimilarity zeros.

Table S18. Correlation coefficients for 21 plant community properties for the German dataset.

Table S19. Correlation coefficients for 21 plant community properties for the US dataset.

Table S20. Variance explained by 12 PCA axes (12 vif-selected community properties).

Table S21. PCA scores for 12 vif-selected community properties of PCA's in Fig. 1.

Table S22. Full dataset of community properties for all plots used in the PCA's over all years (submitted along with R-code at submission).